Retail Credit Modeling

Wei Cai, Miao Pan, Zeyu Xiong, Lipeng Yu, Yingchi Chen

1.1 PURPOSE

Provide detailed description of Small Business Behavioral Score, provide details about model assumption, data preprocessing, model limitation, model assessment and scorecard development.

1.2 PORTFOLIO

The portfolio consists of startup, term loan, demand loan, visa, OLL and widely held customer

1.3 MODEL USE

To help predict the exposure to lots of small businesses so that the institution can make a good balance between risk and reward.

1.4 MARKET OUTLOOK

Small business loans constitute more than a quarter of the lending volume in the US, it's playing a more and more important role in retail lending. The model helps to quickly decide whether the bank should lend money to the other to enlarge bank's gains. So a more accurate and efficient model can help bank manage the credit risk better.

1.5 MODEL DEVELOPMENT PROCESS

Determine business objectives → Data Preparation → Model development → Model assessment/approval

→ Model deployment → Model Monitoring

2.1 DATA SOURCE

Data consists of Business Bureau variables, Customer Bureau variables, application data, customer relation data and loan performance data.

2.2 TIMEFRAMES

In the dataset, there are 4 observation point --- Jan 2014, Apr 2014, July 2014, Oct 2014, so for each observation point, 24 months before it is observation period, and 12 months after it is performance period.

2.3 TARGET VAIRABLE DEFINITION

I use t12 as target variable meaning default flag in 12 months after the observation point, in other word, the model will be used to predict whether it will default in 12 months.

2.4 POPULATION EXCLUSIONS

There are 5 widely help customers and 11 deceased customers, it's important to exclude deceased customers because they won't default anymore which is no use for building the model.
There are 9028 customers, and 16 of them are excluded and is a pretty minor part, so it doesn't affect anything.

2.5 MODELING POPULATION

There are only 900 customers are in default, so the default rate is 9.986% for total population

2.6 EXPLANATORY VARIABLES

I use debit in each month divided by the credit in each month to get the ratio in each month---from current month to previous 12 months, in this way, I could know the percentage of debits taken by credits.

2.7 SEGMENTATION

We can segment this population based on customer type—whether it's startup, it's term loan customer or demand loan customer. Alternatively, we can segment them using industry type—doctor, restaurant…… And we can use time key to segment the population so that we can build model for each observation point. For our group, we don't segment the population so we only need to build one model.

2.8 SAMPLING METHODOLOGY

Based on the time key, we decide to all data observed at Oct 2014 as out of time validation samples, and we use the rest data to train and cross-validate our model which makes more sense, as we always need to predict future using model built on previous data.
The default rate in test set is 10.21%, and the rate is 9.91% in train set, both are very close, so we are sure that even though the target variable in each sample is biased, but the sampling method is not biased.

3.1 MODELING CONSIDERATIONS

Modeling technique is appropriate because the dataset has a large number of observations and variables, it's difficult to assess the quality of an observation by human, so we need to turn to the help of some prediction model, and regression is the most frequently used one.

3.2 VARIABLE REDUCTION

3.2.1 Pre-Screen

After excluding t1 – t12 which will be target variable, there are 561 – 12 = 549 explanatory variables ‘widely held customer’, ‘deceased’ should be excluded because it has only one value if we exclude all decreased and widely held customer.

And for business consideration, we should exclude ‘NFP’-not for profit variable, because the business is not for profit, it's no use to predict whether it will default or not.

3.2.2 Univariate Screening

We use Decision Tree Classifier to bin each variable into 10 bins, and by computing the WOE in each bin, we summarize the IV of debt/credit ratio in previous months as follows

```
--ratio Previous--0     0.0903
--ratio Previous--1     0.0969
--ratio Previous--2     0.0842
--ratio Previous--3     0.0827
--ratio Previous--4     0.1207
--ratio Previous--5     0.0894
--ratio Previous--6     0.1544
--ratio Previous--7     0.114
--ratio Previous--8     0.0861
--ratio Previous--9     0.1114
--ratio Previous--10    0.1001
--ratio Previous--11    0.0722
--ratio Previous--12    0.1037
```

As we can see, the IV of these ratio are very similar, they are all on the margin between medium predictive power and lower predictive power.

Given those variables with WOE as information, we compute their IV as follows

WOE_ALL2320	0.38734141828162394
WOE_ALL2326	0.29322107333589087
WOE_ALL2327	0.42405199827192636
WOE_ALL2350	0.29893666381580375
WOE_ALL2358	0.2602323945943047
WOE_ALL2380	0.2002920516525641
WOE_ALL2700	0.18305346657633959
WOE_ALL6160	0.19519973617863415
WOE_ALL6200	0.3273335468876101
WOE_ALL6230	0.1342800479492242
WOE_ALL7330	0.37000425472009046
WOE_ALL7938	0.3826940198895168
WOE_ALL8150	0.3192846506053846
WOE_ALL8160	0.209381321601808
WOE_ALL8358	0.24603887793902038
WOE_BCA2358	0.18012682425735535
WOE_BCA2380	0.1315597839423258
WOE_BCA5030	0.26480781507020545
WOE_BCC3510	0.13341395708283696
WOE_BCC3515	0.22270264248071991
WOE_BCC5620	0.3485676273746696
WOE_BCC5830	0.34385352672613245
WOE_BCC6200	0.2589355913827466
WOE_BCC6280	0.25598015066479785
WOE_BCC7110	0.5859092891193196
WOE_BCC7120	0.5834325181583151
WOE_BRC8158	0.1769650981373258

WOE_CVPRAEP112	0.12160411618807856
WOE_CVPRAGG102	0.10783248256622288
WOE_CVPRAGG501	0.28528539726636754
WOE_CVPRAGG512	0.18562965019958416
WOE_CVPRAGG519	0.17371133243952336
WOE_CVPRAGG905	0.5669307020356739
WOE_CVPRAGG907	0.24232821542840072
WOE_CVPRAGG910	0.5617639984653171
WOE_CVPRRVLR01	0.4883909963620743
WOE_CVPRRVLR07	0.40470260818509207
WOE_CVPRTPR103	0.1891794652412278
WOE_CVPRTPR212	0.30611117390035836
WOE_CVPRTPR301	0.14904875712901874
WOE_CVPRTPR312	0.2170372998947847
WOE_CVPRTRV04	0.20713823180569857
WOE_CVPRTRV12	0.11963430576511834
WOE_CVPRTRV14	0.16762914994798317
WOE_CVPRWALSHR01	0.164193186
WOE_CVPRWALSHR02	0.143961182
WOE_CVSC100	0.978163231147247
WOE_CVSC110	0.8468111338079767
WOE_GO170	0.8740953663475772
WOE_HLC5030	0.17029937459511502
WOE_HLC7110	0.28231108787466797
WOE_IQT9410	0.14458337140551775
WOE_IQT9420	0.16080695977119985
WOE_NA11	0.6866740795846351

WOE_PD_Total_Scorecard_Points	1.78
WOE_REV2320	0.308208451682837
WOE_REV2327	0.28303867682232087
WOE_REV2328	0.33441540385278434
WOE_REV2350	0.17813261319039447
WOE_REV3423	0.36107832163880343
WOE_REV5020	0.11441918266324785
WOE_REV5030	0.1221078035856098
WOE_REV5620	0.40593758882156317
WOE_REV8153	0.13034709576380016
WOE_TBSAT103S	0.2405215161393096
WOE_TBSAT33A	0.4128352491124638
WOE_TBSAT34B	0.5427210878099359
WOE_TBSBC104S	0.3295775639658942
WOE_TBSBC33S	0.1324118359194346
WOE_TBSBC35S	0.14581218479673902
WOE_TBSBC97A	0.2494620382254372
WOE_TBSBR34S	0.6097007482428698
WOE_TBSG001B	0.7055987734684731
WOE_TBSG059S	0.7047688600663347
WOE_TBSG202A	0.2987950891673159
WOE_TBSG302S	0.6249062952372191
WOE_TBSRE24S	0.11410747403043064
WOE_TBSRE29S	0.36349635088476984
WOE_TBSRE30S	0.4554595703635355
WOE_TBSRE36S	0.551873464754903
WOE_TBSRN33S	0.3578667987129126
WOE_TBSRN34S	0.4277421767488725
WOE_TBSSC100	0.6677551766514037

WOE_cust_rev_max_dlq_6mos	1.154
WOE_cust_sum_dlq_24mos	1.18978746277
WOE_dda_av_bal	0.5015145212465746
WOE_dda_avg_dly_dep_amt_L90	0.394
WOE_dda_max_Avg_Cr_Bal	0.95325970563
WOE_dda_min_Min_Mthly_Bal	1.007
WOE_dda_min_avail_bal	0.41179595796
WOE_dda_sum_Acc_Db_Bal	0.70992536131
WOE_dda_sum_OS_Bal	1.13075358451
WOE_dda_sum_Ttl_Dep_Prev	0.424
WOE_max_ks_age	0.10723118087259698
WOE_max_ks_max_dlqdays_6mos	0.456
WOE_max_ks_max_util_3mos	0.624
WOE_max_ks_max_util_6mos	0.516
WOE_max_ks_num_dlqdays_12mos	0.435

Clearly, they are variables with very strong predictive power with the lowest one equals to 0.107

Firstly, I delete those variables with more than 99.5% of nan, so there are 286 variables left waiting to be compute IV. And after computing their information values, we find that there are 104 variables with IV less than 0.1, so 104 variables are filtered out.

In particular, we pick ‘SPP_Group_1’ variable with IV = 0.005, its results are following

	Default Number	Non-default	Default Rate	WOE	IV
N	818	7201	10.2%	-0.0236	0.0005
Y	82	911	8.26%	0.209	0.0044
Total	900	8112	9.98%		0.0049

And we pick 'TBSSC100' variable with IV = 1.28

	Default Number	Non-default	Default Rate	WOE	IV
Bin1	227	2259	9.13%	0.099	0.0026
Bin2	50	354	12.37%	-0.241	0.0029
Bin3	27	1566	1.69%	1.862	0.3
Bin4	268	70	70.58%	-3.074	0.547
Bin5	91	1125	7.48%	0.316	0.012
Bin6	55	1526	3.47%	1.124	0.143
Bin7	139	514	21.28%	-0.891	0.081
Bin8	99	137	41.94%	-1.874	0.174

Bin9	28	467	5.66%	0.615	0.016
Bin10	16	94	14.54%	-0.428	0.0026
Total	900	8112	9,98%		1.28

If the default rate in the bin is higher than the total default rate, than WOE will be negative, and as default rate increases, the magnitude of WOE increases.

3.2.3 Multivariate screening

After excluding those insignificant variables, we do var_clus to the remaining 182 variables, and we get 22 clusters, we just show several of them. As the first two clusters shown below, we can easily find that, generally, variables collected using similar ways will be gathered at the same cluster. Besides, variables collected by different ways can be grouped into the same cluster.
Following the course, we pick the one with the highest R-square with its own cluster component and another one with the highest information value in each cluster. So we will select 44 variables.

```
--cluster-0-0-0-0-0
  |-----ALL2320
  |-----ALL2350
  |-----ALL2380
  |-----ALL2700
  |-----ALL7330
  |-----ALL8150
  |-----ALL8358
  |-----BCA2380
  |-----BCC5620
  |-----BRC8158
  |-----CVPRAEP112
  |-----CVPRAGG519
  |-----CVPRAGG905
  |-----CVPRAGG907
  |-----CVPRAGG910
  |-----CVPRRVLR01
  |-----CVPRTPR103
  |-----CVPRTRV12
  |-----CVPRWALSHR02
  |-----GO170
  |-----REV2320
  |-----REV2350
  |-----REV5030
```

```
-cluster-0-1-0-0-0-0-1-1
 |-----max_ks_any_bad_12mos
 |-----max_ks_any_bad_24mos
 |-----max_ks_any_bad_3mos
 |-----max_ks_any_bad_6mos
 |-----max_ks_num_bad_12mos
 |-----max_ks_num_bad_24mos
 |-----max_ks_num_bad_3mos
 |-----max_ks_num_bad_6mos
```

```
-cluster-0-0-0-1
 |-----ALL2326
 |-----ALL2327
 |-----ALL2358
 |-----BCA2358
 |-----CVPRTPR312
 |-----NA11
 |-----REV2327
 |-----REV2328
```

```
-cluster-0-1-0-0-0-1-0
 |-----TBSAT103S
 |-----TBSG001B
 |-----TBSG059S
 |-----TBSG302S
 |-----TBSRE24S
 |-----TBSRE36S
 |-----credit_prev7
 |-----debit_curr
```

3.3 Model Fitting

For simplicity, we decide to use 182 variables to select 44 variables, variables we select are following

```
TBSRE36S
GO170
CVPRAEP112
CVPRAGG501
CVPRWALSHR01
CVPRWALSHR02
cust_max_dlq_3mos
dda_max_open_date
dda_max_Last_Dep_Date
dda_sum_Acc_Db_Bal
dda_sum_Ttl_Dep_Prev
min_oll_avg_ytd_bal
max_oll__days_late
tot_oll_os_bal
max_oll_num_dlq_12mos
max_oll_term_max_dlq_12mos
max_oll_max_dlqdays_3mos
ks_tot_limit
max_ks_any_bad_12mos
max_ks_num_bad_6mos
max_ks_max_dlq_24mos
max_ks_max_dlq_12mos
max_ks_age
```

```
ALL2350
BCC3510
BRC8158
HLC5030
REV2350
REV5020
REV5620
debit_curr
debit_prev6
debit_prev9
credit_curr
credit_prev2
credit_prev7
credit_prev8
credit_prev10
ratio_prev4
ratio_prev6
ratio_prev7
ratio_prev9
ratio_prev10
ratio_prev12
```

Most variables are from Business Bureau, Customer Bureau variables and loan performance.

3.4 Scorecard Scaling

According to the model we built, we can predict the probability of default or non-default for each observation, so we can calculate the odds for each observation, therefore, we are able to give the score for each observation directly.

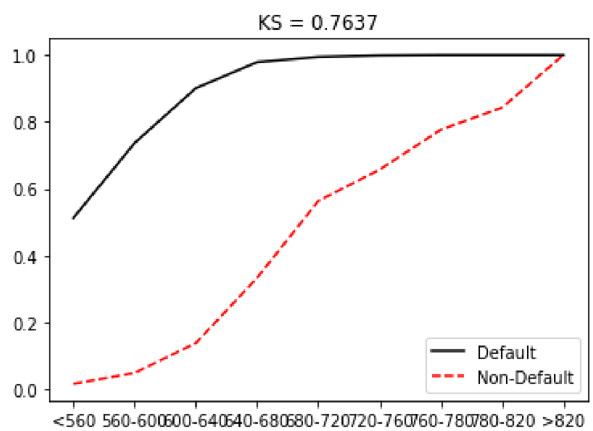
For example, for the first observation in the sample, our model predict odds(non-default : default) = 0.792/0.208 = 3.81. so its final score $= 633.56 + 28.8539 * \log 3.81 = 672$

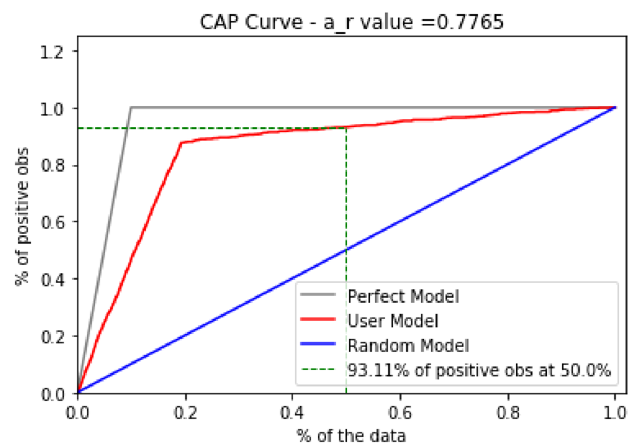
So the score of first 10 observation are 672, 844, 752, 654,748, 690, 416, 492, 627, 614, the score of 7th and 8th observation is low which is consistent with the target variable in which they are all default.

3.5 SCORECARD ASSESSMENT

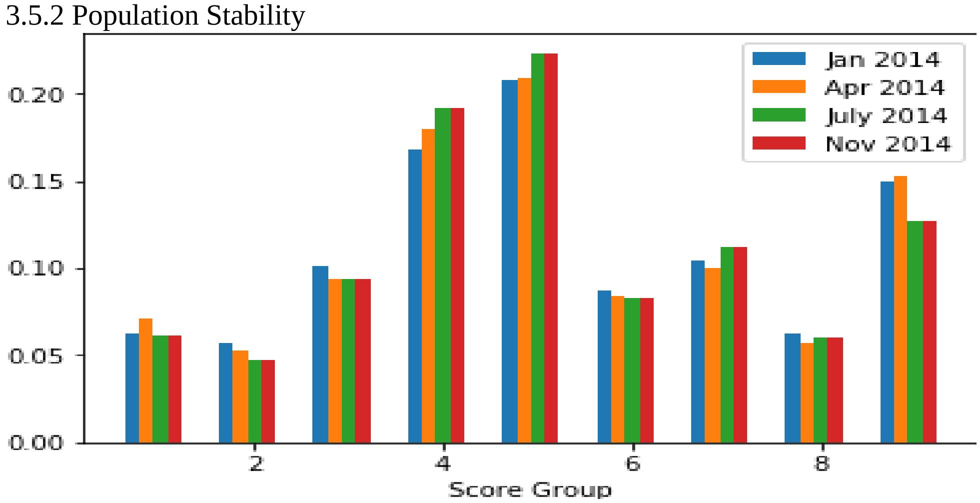
3.5.1 Rank-Ordering

We only build one model for all samples, two cumulative distributions figures are shown below, from which we calculate KS = 76.37%, and AR = 77.65%





3.5.2 Population Stability



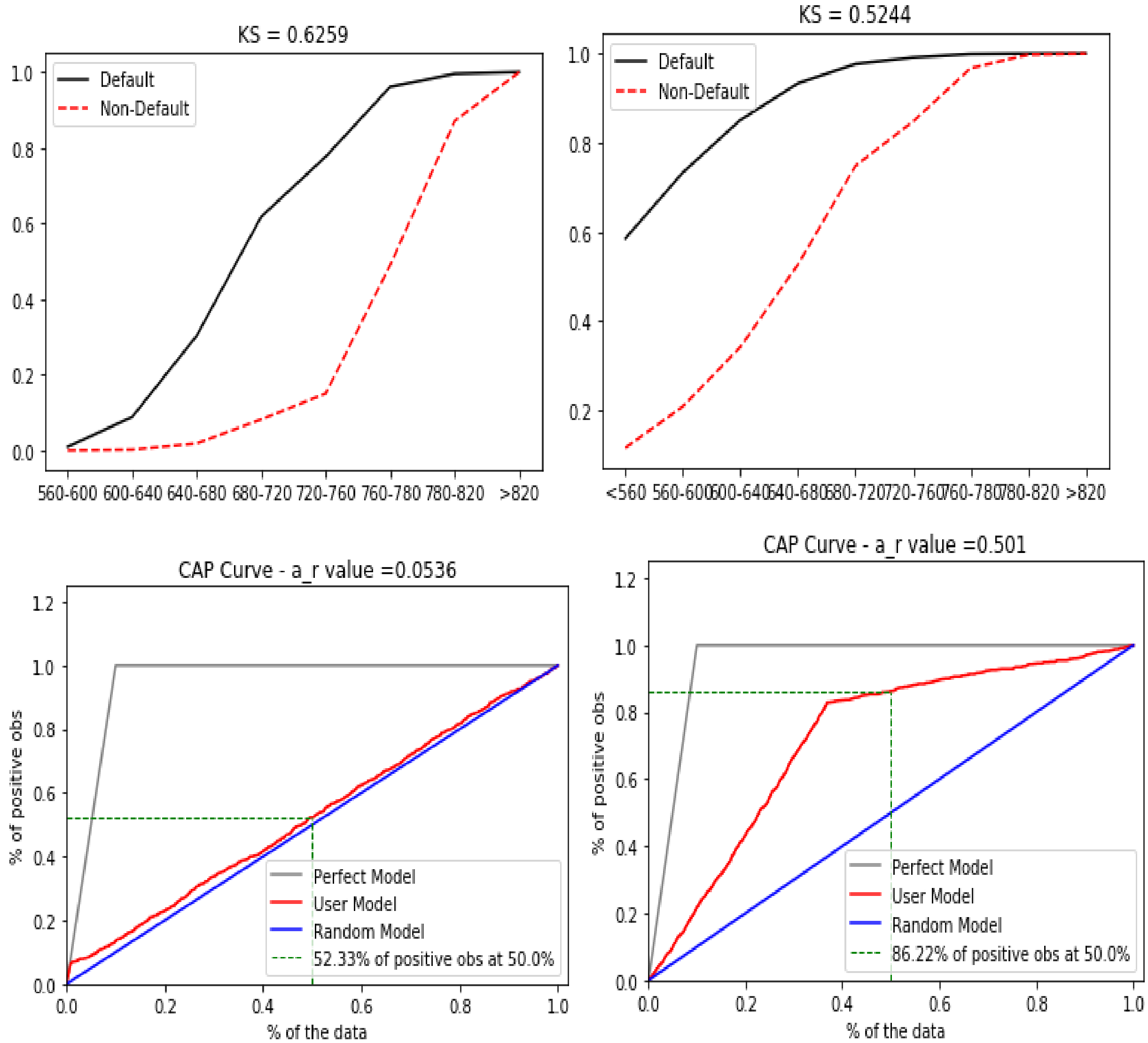
We also compute PSI which are all less than 0.1, so there is no significant shift in population.

3.5.3

Since there are nan in benchmark2 and benchmark3, we only use 1 and 4 to do comparison.

As the following figures show, even though the KS statistic of benchmark1 is larger than benchmark4, CAP curve shows that benchmark4 is much better than benchmark1, where benchmark1 is only slightly better than a random model.

Luckily, our model is better than both benchmarks in terms of KS and CAP curve



4.0 MODEL LIMITATIONS AND ASSUMPTIONS

- It assumes that the population feature won't change, and the future will follow the rule of the past.
- By excluding variables in part 3, we may lose a part of information which may be useful for prediction.